The Cortical Conductor Theory: Towards Addressing Consciousness in AI Models

Joscha Bach (joscha@bach.ai)
Harvard Program for Evolutionary Dynamics, One Brattle Square #6
Cambridge, MA 02139 USA

Abstract

AI models of the mind rarely discuss the so called "hard problem" of consciousness. Here, I will sketch informally a possible functional explanation for phenomenal consciousness: the conductor theory of consciousness (CTC). Unlike IIT, CTC is a functionalist model of consciousness, with similarity to other functionalist approaches, such as the ones suggested by Dennett and Graziano.

Keywords: phenomenal consciousness, cortical conductor theory, attention, executive function, binding, IIT

"No computer has ever been designed that is ever aware of what it's doing; but most of the time, we aren't either."
– Marvin Minsky

Introduction: Artificial Intelligence as a computational science of the mind

Understanding the nature of our minds and their relationship to the universe has always been one of the most significant questions philosophy sought to address.

For centuries, scientists and philosophers emphasized the role of mathematics in this quest. The discovery of the idea of computation and its formalization in the 1920ies by Alan Turing and Alonzo Church paved the way to a new way of thinking about thinking, and replaced the old intuitions of mechanistic philosophy with more precise ones of computationalism, and opened up the way of building the new family of theories of computational systems.

The relationship between mathematics and computation is not trivial, and even though computation is defined mathematically (and constructive mathematics is arguably computational), it makes sense to understand them as separate realms. Mathematics is the domain of all formal languages, and allows the expression of arbitrary statements (most of which are uncomputable). Computation may be understood in terms of computational systems, for instance via defining states (which are sets of discernible differences, i.e. bits), and transition functions that let us derive new states. Whereas mathematics is the realm of specification, computation is the realm of implementation; it captures all those systems that can actually be realized.

Computational systems are machines that can be described apriori and systematically, and implemented on every substrate that elicits the causal properties that are necessary to capture the respective states and transition functions.

The absence of an understanding of substrate independent machines lead Leibniz to the rejection of mechanist philosophy: *"Perception, and what depends on it i.e., cognition], is inexplicable in a mechanical way, that is, using figures and motions. Suppose there would be a machine, so arranged as to bring forth thoughts, experiences and perceptions; it would then certainly be possible to imagine it to be proportionally enlarged, in such a way as to allow entering it, like into a mill. This presupposed, one will not find anything upon its examination besides individual parts, pushing each other—and never anything by which a perception could be explained."* (1714). Conversely, its inkling prompted Julien Offray de LaMettrie's (1748) remark that while minds are machines, these had to be thought of as *"immortal"* and *"transcendental"*.

Computation is sometimes seen in opposition to dynamical systems (see van Gelder 1998), and we can distinguish different classes of computational systems to account for that, based on the classes of functions they can compute effectively (in the unlimited case) and efficiently (with reasonably bounded resources), starting from (deterministic or probabilistic) finite state machines over Turing Machines to different classes of hyper-computers capable of continuous state change or even a-causal computers that may allow a transition function to use information from a future state of the machine. We find that while dynamical systems often cannot be effectively computed on a finite state machine (such as a von Neumann computer), they can often be efficiently approximated. (The metaphysical implications of whether our universe can only realize finite state machines or hyper-computation are profound and sometimes of concern in the philosophy of mind, but outside the scope of this discussion.)

The formation of a new, computational study of the mind was fraught with difficulty from the start. By the 1950ies, the influence of positivism had lead to the emergence and entrenchment of behaviorism in psychology, which stifled theoretical psychology and made it evidently impossible for psychologists to formulate comprehensive theories of the mind, so a new discipline was established: Artificial Intelligence was the attempt of thinkers like Marvin Minsky, John McCarthy and others to treat the mind as a computational system, and thereby open its study to experimental exploration by building computational machines that would attempt to replicate the functionality of minds.

Artificial Intelligence soon formed two camps: one that was dedicated to the study of intelligence, and one that focused on the automation of tasks that required human intelligence. While both camps developed applications and theories and often worked on similar systems, the rift between "cognitive AI " and "narrow AI" widened, partly because large factions of the cognitive AI camp championed symbolic approaches and rejected neural learning as simplistic. The failure to deliver on some of the early, optimistic promises of machine intelligence, as well as cultural opposition, lead to cuts in funding for cognitive AI, and eventually the start of the new discipline of Cognitive Science. However, Cognitive Science did not develop a cohesive methodology and theoretical outlook, and became an umbrella term for neuroscience, AI, cognitive psychology, linguistics and philosophy of mind.

In the last five years, AI research has been dominated by the success of deep learning, which was fueled by theoretical insights into the training of neural networks with a large number of hidden layers, advances in computer hardware, and partially by the availability of large amounts of training data. The rapid advances of learning machines have lead to a renewed interest in the original goals of AI, as well as the dissemination and development of ideas on the nature of learning, perception, and mental representation. However, the recent progress was arguably driven by successes in the narrow AI camp, and AI as a field is not very much concerned with the study of minds any more. Progress on this front will likely require a better understanding of our mental architecture, reasoning, language, reflection, self model and consciousness.

Consciousness in cognitive science

While AI offers a large body of work on agency, autonomy, motivation and affect, cognitive architectures and cognitive modeling, there is little agreement on how to address what is usually called "the hard problem" of consciousness. How is it possible that a system can take a first person perspective, and have phenomenal experience?

One of the better known recent attempts to address phenomenal consciousness is Guilio Tononi's Integrated Information Theory (IIT) (2012, 2016), which has been championed by the neuroscientist Christof Koch and the physicist Max Tegmark (2014). Perhaps not entirely unlike Leibniz, Tononi argues that experience cannot be reduced to a functional mechanism, and hence it must be an intrinsic property of a system, rather than a functional one. He characterizes consciousness by a parameter, Φ, which is a measure for the amount of mutual information over all possible partitionings of an information processing system. If the information in the system is highly integrated (i.e. the information in each part of the system is strongly correlated with the information in the others), it indicates a high degree of consciousness. As has for instance been argued by Aaronson (2015), IIT's criterion of information integration could perhaps be necessary, but is not sufficient, because we can construct structurally trivial information processing systems that maximize Φ by maximally distributing information (for instance via highly interconnected XOR gates). Should we assign consciousness to processing circuits that are incapable of exhibiting any of the interesting behaviors of systems that we usually suspect to be conscious, such as humans and other higher animals?

From a computationalist perspective, IIT is problematic, because it suggests that two systems that compute the same function by undergoing a functionally identical sequence of states might have different degrees of consciousness based on the arrangement of the computational elements that realize the causal structure of the system. A computational system might turn out to be conscious or unconscious regardless of its behavior (including all its utterances professing its phenomenal experience) depending on the physical layout of its substrate, or the introduction of a distributed virtual machine layer.

A more practical criticism stems from observing conscious and unconscious people: a somnambulist (who is generally not regarded as conscious) can often answer questions, navigate a house, open doors etc., and hence should have cortical activity that is distributed in a similar way as it is in an awake, conscious person (Zadra et al. 2013). In this sense, there is probably only a low quantitative difference in Φ, but a large qualitative difference in consciousness. This qualitative difference can probably be explained by the absence of very particular, local functionality in the brain of the somnambulist: while her cortex still produces the usual content, i.e. processes sensory data and generates dynamic experiences of sounds, patterns, objects, spaces etc. from them, the part that normally attends to that experience and integrates it into a protocol is offline. This integrated experience is not the same as information integration in IIT. Rather, it is better understood as a particular local protocol by one of the many members of the "cortical orchestra" : its *conductor*.

In this contribution, I will sketch how a computational model can account for the phenomenology and functionality of consciousness, based on my earlier work in the area of cognitive architectures (Bach 2009); we might call this approach the "conductor theory of consciousness" (CTC).

An AI perspective on the mind

Organisms evolved information processing capabilities to support the regulation of their systemic demands in the face of the disturbances by the environment. The simplest regulator system is the feedback loop: a system that measures some current value and exerts a control operation that brings it close to a target value. Using a second feedback loop to regulate the first, the system can store a state and regulate one value depending on another. By changing the control variables to maximize a measured reward variable, a system can learn to approximate a complex control function that maps the values of a set of inputs (sensors) to operators (effectors).

Our nervous systems possess a multitude of feedback loops (such as the mechanisms of the brain stem regulating heart rate and breathing patterns).

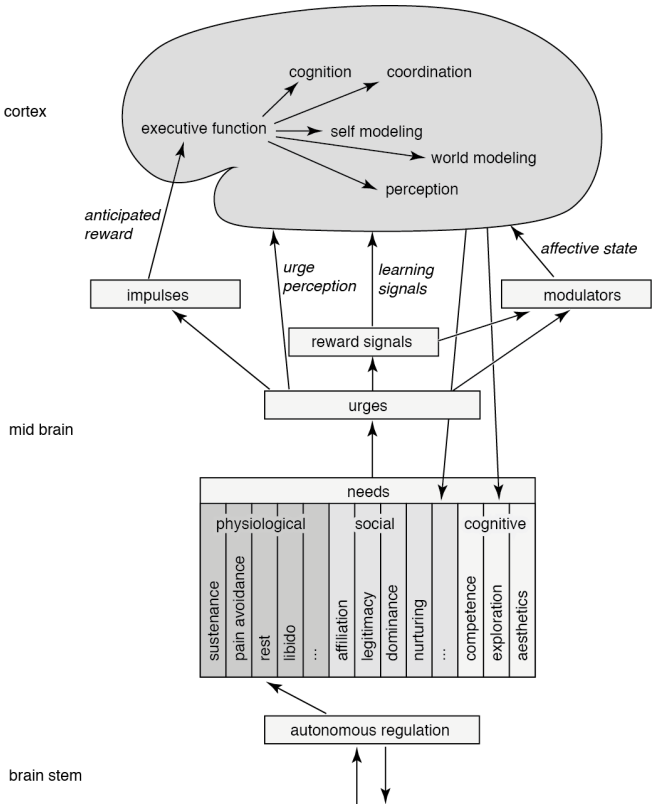

Figure 1. From needs to cognition

The control of behavior requires more complex signals; the sensors of the limbic system measure changes in organismic demands and respond to satisfaction of needs with pleasure signals (indicating to intensify the current behavior). The frustration of needs leads to displeasure signals (pain) which indicate the current behavior should be stopped.

Directed behavior of a system may be governed by impulses, which associate situations (complex patterns in the outer or inner environment of the organism) with behavior to obtain future pleasure, or avoid future pain. Pain and pleasure act as reward signals that establish an association between situations, actions and needs (see Bach 2015). In mammals, such connections are for instance established in the hippocampus (see for instance Cer and O'Reilly 2006).

The human neocortex enables better regulation of needs by encoding sensory patterns into a complex hierarchical model of the environment (including the inner environment). This dynamic model is not just a mapping from past observation to future observations, but takes on the shape of a progressively updated stateful function, a program that generates a simulation of the environment.

The formation of the model is driven largely by data compression, i.e. by optimizing for the data structure that allows the best predictions of future observations, based on past observations. This principle has for instance been described by Ray Solomonoff (1964): The best possible model that a computational agent can form about its environment is the shortest program among those that best predict an observation from past observations, for all observations and past observations.

Machine learning models of the mind can be understood as approximating Solomonoff induction (see Hutter 2005), by capturing the apparent invariances of the world into an almost static model, and its variance as a variable state of that model. By varying the state, such a model cannot only capture the current state of the world, but be used to anticipate and explore possible worlds, to imagine, create and remember. Machine learning systems have demonstrated how recurrent neural networks can discover and predict the structure of visual and auditory stimuli by forming low level feature detectors, which can then be successively organized into complex high level features, object categories and conceptual manifolds (LeCun, Bengio, Hinton 2015). Deep networks can form hierarchical knowledge representations. LSTMs (Hochreiter and Schmidhuber 1997) and GRUs (Cho et al. 2014) are building blocks for recurrent neural networks that can learn sequences of operations. Generative neural networks can use the constraints learned from the data to produce possible worlds (Dosovitskiy et al. 2015).

While current machine learning systems outperform humans in many complex tasks that require the discovery and manipulation of causal structures in large problem spaces, they are very far from being good models of intelligence. Part of this is due to our current learning paradigms, which lead to limitations in the generation of compositional knowledge and sequential control structures, and will be overcome with incremental progress. Recently, various researchers have proposed to introduce a unit of organization similar to cortical columns into neural learning (Hinton et al. 2011). Cortical columns are elementary circuits containing between 100 and 400 neurons (Mountcastle 1997), and are possibly trained as echo state networks (Jaeger 2007) to achieve functionality for function approximation, conditional binding and reward distribution. In the human neocortex, the columnar units form highly interconnected structures with their immediate neighbors, and are selectively linked to receptive fields in adjacent cortical areas. A cortical area contains ca. 10^6 to 10^7 columns, and may be thought of as a specialized instrument in the orchestra of the neocortex.

Beyond current machine learning

A more important limitation of many current machine learning paradigms is their exclusive focus on policy learning and classification. Our minds are not classifiers—they are simulators and experiencers. Like machine learning systems, they successively learn to identify features in the patterns of the sensory input, which they then combine into complex features, and organize into maps. High-level features may be integrated into dynamic geometries and objects, motor patterns and procedures, auditory structure and so on. Features, objects and procedures are sensory-motor scripts that allow the manipulation of mental content and the execution of motor actions.

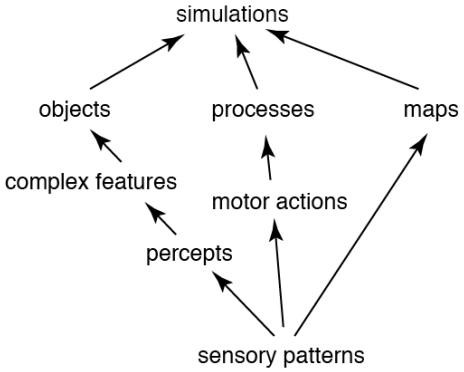

Figure 2. Gradual abstraction from sensory patterns to mental simulations

Unlike most machine learning systems, our minds combine these objects, maps and procedural dynamics into a persistent dynamic simulation, which can be used to continuously predict perceptual patterns at our systemic interface to the environment (figure 2). The processing streams formed by the receptive fields of our cortical instruments enable the bottom-up cuing of perceptual hypotheses (objects, situations etc.), and trigger the top-down verification of these hypotheses, and the binding of the features into a cohesive model state.

The elements of this simulation do not necessarily correspond to actual objects in the universe: they are statistical regularities that our mind discovered in the patterns at its systemic interface. Our experience is not directed on the pattern generator that is the universe, but on the simulation produced in our neocortex. Thus, our minds cannot experience and operate in an "outer" reality, but in a dream that is constrained by the available sensory input and the context of previous input (Bach 2011).

Human cognition does not stop at generative simulations, however. We can abstract our mental representations into a conceptual manifold (figure 3). Concepts can be thought of as an address space for our sensory-motor scripts, and they allow the interpolation between objects, as well as the manipulation and generation of previously unknown objects via inference. The conceptual manifold can be organized and manipulated using grammatical language, which allows the synchronization of concepts between speakers, even in the absence of corresponding sensory-motor scripts. (The fact that language is sufficient to infer the shape of the conceptual manifold explains the success of machine translation based on the statistical properties of large text corpora, despite the inability of these systems to produce corresponding mental simulations.)

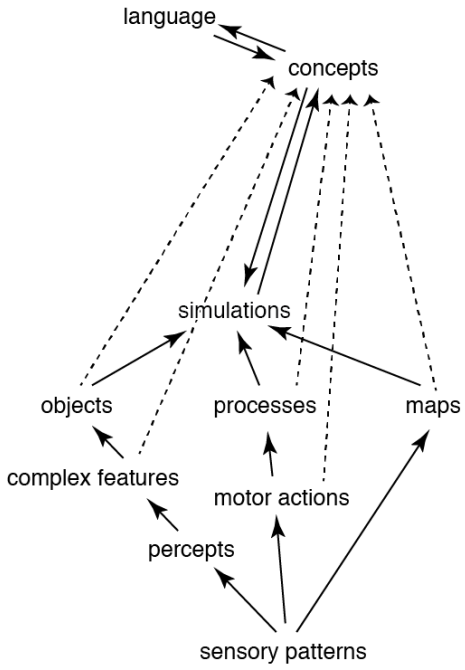

Figure 3: Conceptual abstraction

The cortical conductor

Cortical columns may be thought of as elementary agents that self-organize into the larger organizational units of the brain areas as a result of developmental reinforcement learning. The activity of the cortical orchestra is highly distributed and parallelized, and cannot be experienced as a whole. However, its performance is coordinated by a set of brain areas that act as a conductor. The conductor is not a "homunculus", but like the other instruments, a set of dynamic function approximators. Whereas most cortical instruments regulate the dynamics and interaction of the organism with the environment (or anticipated, reflected and hypothetical environments), the conductor regulates the dynamics of the orchestra itself. Based on signals of the motivational system, it provides executive function (i.e. determines what goals the system commits to at any given moment), resolves conflicts between cortical agents, and regulates their activation level and parameterization. Without the presence of the conductor, our brain can still perform most of its functions, but we are sleep walkers, capable of coordinated perceptual and motor action, but without central coherence and reflection.

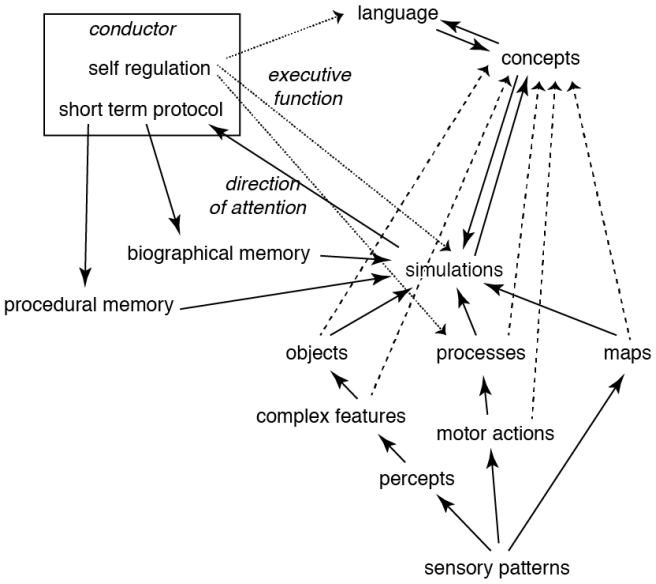

Figure 4: The cortical conductor

In the human brain, the functionality of the conductor is likely facilitated via the dorsolateral prefrontal cortex (Bodovitz 2008, Safavi 2014, Del Cul 2009), anterior cingulate cortex and anterior insula (Fischer et al. 2016). The conductor has attentional links into most regions. In each moment, it directs its attention to one or a few of the cortical instruments, while others continue to play unobserved in the background. The conductor may not access the activity of the region it attends to in its entirety, but it may usually access some of the currently relevant processing states and parameters of it.

To learn and to reflect, the conductor maintains a protocol of what it attended to, as a series of links to experiences generated by the other cortical instruments. This protocol may be used to address the currently active regions, and to partially recreate past states of the mental simulation by reactivating the corresponding configuration of active regions with the parameters of the stored links. The reactivation of a past state of the mental simulation will generate a re-enactment of a previous world state: a memory. Further abstraction of the protocol memory leads to the formation of new kinds of sensory motor scripts: an autobiographical memory (events that happened to the agent), and a procedural memory.

The reflective access to the protocol allows learning and extrapolation of past events, and the act of accessing the protocol may of course itself become part of the protocol. By accessing the memory of the access to its own protocol, the system remembers having had access to experience (access consciousness).

While all cortical regions store information as a result of updating their models and learning associations to motivational signals, the attentional protocol of the conductor is the only place where experience is integrated. Information that is not integrated in the protocol cannot become functionally relevant to the reflection of the system, to the production of its utterances, the generation of a cohesive self model, and it cannot become the object of access consciousness.

Phenomenal consciousness may simply be understood as the most recent memory of what our prefrontal cortex attended to. Thus, conscious experience is not an experience of being in the world, or in an inner space, but a memory. It is the reconstruction of a dream generated more than fifty brain areas, reflected in the protocol of a single region. By directing attention on its own protocol, the conductor can store and recreate a memory of its own experience of being conscious.

The idea that we are not actually conscious in the moment, but merely remember having been conscious is congruent with known inconsistencies in our experience of consciousness, such as subjective time dilation, false continuity, and loops in the conscious experience.

Subjective dilation of time results from states of high arousal, for instance during an accident, whereas uneventful flow states often lead to a subjective contraction of time. Both dilated and contracted time do not correspond to an increase or decrease in the actual processing speed of our cognitive operations. Instead, they result from a higher or lower number of entries in the protocol memory: the experienced time interval only seems to be longer or shorter with hindsight. An extreme case of a subjective dilation of time can happen during dreams, which sometimes play out in a physical time interval of a few seconds of REM sleep, yet may span hours of subjective time. This may be explained by the spontaneous generation of the entire dream, rather than the successive experience of each event. Hour-long dreams are probably simply false memories.

False continuity results from gaps in our attention, for instance during saccadic movements, or interruptions and distractions of gaze. While these breaks in attention may lead to missing significant changes in parts of the environment that we believe we are attending to, they are not part of the protocol memory and hence our experience appears to be unbroken in hindsight. For a considerable fraction of our days, we are probably wakeful but not conscious.

Inconsistent experiences of consciousness can be explained as false memories, but they do not have subjective qualities that makes them appear “less conscious” than consistent experiences. Thus, if at least some of our conscious experience is a false memory, why not all of it?

Treating consciousness as a memory instead of an actual sense of the present resolves much of the difficulty for specifying an AI implementation of consciousness: it is necessary and sufficient to realize a system that remembers having experienced something, and being able to report on that memory.

Consciousness and self model

In the above discussion, I have treated phenomenal consciousness in the sense of "the feeling of what it's like". However, consciousness is often associated with more concrete functionality, especially a specific model of self, and a set of functionality pertinent to that model. This has lead Marvin Minsky (2006) to call consciousness "a suitcase term", a notion that is notoriously hard to unpack.

Conscious states differ by the configuration and available functionality of a cognitive system at a given time. However, once we understand how an attentional protocol can provide for binding of other cortical functionality into a single structure for the purpose of self regulation, we can enumerate some of the functionality that corresponds to a given conscious state.

Core consciousness is characterized by:
- a local perceptual space
- the ability to access mentally represented percepts
- a current world model
- directed attention (inwards/outwards, wide/focused)
- the ability to access and follow concepts and similar content
- the ability to manipulate and create concepts and similar content,
- the presence of an inner stage of currently active, non-perceptual concepts and associative representations

In deep meditation, the following functionality may be absent:
- an integrated personal self-model (sense of identity)
- a sense of one's own location and perspective in space
- proprioception (position and state of body and limbs)
- valences (pleasure and displeasure signals)
- goals and committed plans
- the awareness of the current affective state
- the influence of desires and urges on behavior
- the ability to create and process discourse (i.e. translate mental representations into communicable symbols, and vice versa)

Lucid dreams are specific dream states that are different from wakefulness by the absence of:
- access to needs/desires, urges
- access to sensory perception
- the ability to exert voluntary control over muscles
- a biographical memory and protocol
- a short term biography
- the ability to separate perceptually grounded content from ideas/imaginations

Dreams are usually in addition characterized by the absence of:
- having accessible knowledge about the fact that there access to percepts and concepts (access consciousness)
- a social model of self (self-ascription of beliefs, desires, intentions, skills, traits, abilities, personality)
- the formation of and access to expectations of immediate future
- the ability to influence behavior based on discursive thought
- the ability to relate self-ascribed actions to apparent mental causes (sense of agency)
- the ability to form memories of the current content
- the ability to reason
- the ability to construct plans
- the ability to act on plans
- the ability to keep goals stable until they are achieved
- the ability to let go of goals that are unattainable
- the ability to signal aspects of one's mental state to others

Diminished states of consciousness (for instance, in small children or due to neurodegenerative diseases) may also impair:
- the ability to influence behavior based on past experience (learning)
- the ability to construct causal models of the environment
- the ability to construct intentional models of agents

The above functionality is part of the general functionality of a human-like cognitive agent and has to be implemented into its cognitive architecture, either explicitly or via a self-organized process of reward driven learning (each of them can be realized on computational machines). The differences between conscious states result from the dissociation or impairment of these functions.

Is the conductor a learned or a predefined structure? I suspect that the formation of the conductor functionality is itself a process of developmental learning, driven by rewards for the self-regulation of cognition, and developmental cues that regulate the onset and some of the parameters of the formation of the structure. Multiple personality disorder lends further credibility to the hypothesis that the conductor is constructed by reward driven neural self-organization. In patients with multiple personalities, the different personas usually do not share a subjective protocol, biographical and procedural memory. But even if we form multiple conductors, they share infrastructure (such as linguistic processing, access to the attentional network and information transfer via the thalamic loop), which ensures that only one of them may be online and form memories at any given moment.

Summary

The cortical conductor theory (CTC) posits that cortical structures are the result of reward driven learning, based on signals of the motivational system, and the structure of the data that is being learned. The conductor is a computational structure that is trained to regulate the activity of other cortical functionality. It directs attention, provides executive function by changing the activity and parameterization and

rewards of other cortical structures, and integrates aspects of the processes that it attended to into a protocol. This protocol is used for reflection and learning. Memories can be generated by reactivating a cortical configuration via the links and parameters stored at the corresponding point in the protocol. Reflective access to the protocol is a process that can itself be stored in the protocol, and by accessing this, a system may remember having had experiential access.

For phenomenal consciousness, it is necessary and sufficient that a system can access the memory of having had an experience—the actuality of experience itself is irrelevant (and logically not even possible).

CTC explains different conscious states by different functionality bound into the self construct provided by the attentional protocol. The notion of integration is central to CTC, however, integration is used in a very different sense than in Tononi's Integrated Information Theory (IIT). In CTC, integration refers to the availability of information for the same cognitive process, within a causally local structure of an agent. In IIT, integration refers to the degree in which information is distributed within a substrate.

CTC is a functionalist theory, and can be thought of as an extension to Dennett's "multiple drafts" model of consciousness (1991). CTC acknowledges that the actual functionality of perception and cognition is distributed, disjoint and fragmentary, but emphasizes the need to integrate access to this functionality for a module that in turn has access to capabilities for reflection and the formation of utterances (otherwise, there would be no self model and no report of phenomenal experience).

CTC also bears similarity to Michael Graziano's attention schema theory of consciousness. Graziano suggests that just like the body schema models the body of an agent, its attention schema models the activity and shape of its attentional network. While the functionality subsumed under access consciousness, phenomenal consciousness and conscious states, and the required mechanisms are slightly different in CTC, we agree with the role of consciousness for shaping and controlling attention-related mechanisms.

Acknowledgements

This work has been supported by the Harvard Program for Evolutionary Dynamics, the MIT Media Lab and the Jeffrey Epstein Foundation. I am indebted to Katherine Gallagher, Adam Marblestone, and the students of the Future of Artificial Intelligence course at the MIT Media Lab for their contributions in discussions of the topic, as well as to Martin Novak and Joi Ito for their support.

References

Aaronson, S. (2015). Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander). http://www.scottaaronson.com/blog/?p=1799. Retrieved 2017-02-15

Bach, J. (2009). Principles of Synthetic Intelligence. Psi, an architecture of motivated cognition. Oxford University Press.

Bach, J. (2011). No room for the mind. Enactivism and Artificial Intelligence. Proceedings of the Conference of History and Philosphy of Computing, Ghent.

Bach, J. (2015). Modeling Motivation in MicroPsi 2. Artificial General Intelligence, 8th International Conference, AGI 2015, Berlin, Germany: 3-13

Bodovitz, S. (2008). The neural correlate of consciousness. Journal for Theoretical Biology, 2008 Oct 7;254(3):594-8

Cer, D.M., O'Reilly, R.C. (2006). Neural mechanisms of binding in the hippocampus and neocortex: Insights from computational models. H.D. Zimmer, A. Mecklinger & U. Lindenberger (Eds) Binding in Memory, Oxford: Oxford University Press

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder--Decoder for Statistical Machine Translation. in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), October 2014, 1724—1734

Del Cul, A., Dehaene, S., Reyes, P., Bravo, E., Slachevsky, A. (2009): Causal role of prefrontal cortex in the threshold for access to consciousness. Brain (2009) 132 (9): 2531-2540

Dennett, D. C. (1992). Consciousness Explained. Back Bay Books , New York

Dosovitskiy, A., Springenberg, J. T., Brox, T. (2015). Learning to Generate Chairs With Convolutional Neural Networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1538-1546

Fischer, D. B., Boes, A. D., Demertzi, A., Evrard, H. C., Laureys, S., Edlow, B. L., Liu, H., Saper, C. B. ,Pascual-Leone, A., Fox, M. D., Geerling, Joel C. (2016). A human brain network derived from coma-causing brainstem lesions. Neurology. Published online before print November 4, 2016, doi: http:/ / dx. doi. org/ 10. 1212/ WNL. 0000000000003404

Graziano, M. S. A., Webb, T. W. (2014). A Mechanistic Theory of Consciousness. International Journal on Machine Consciousness

Hinton, G. E., Krizhevsky, A., Wang , S. D. (2011). Transforming Auto-encoders. International Conference on Artificial Neural Networks. Springer Berlin Heidelberg

Hochreiter S., Schmidhuber, J. (1997). Long short-term memory. Neural Computation. 9 (8): 1735–1780

Hutter, M. (2005). Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability. EATCS Book, Springer

Jaeger, H. (2007). Echo State Network. Scholarpedia, vol. 2, no. 9, pp. 2330

LaMettrie, J. O. (1748): L'Homme Machine

Leibniz, G. W. L. (1714). Monadologie. R. Zimmermann (ed.), Wien, Draumüller und Seidel 1847

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep Learning. Nature 521, 436–444

Minsky, M. (2006). The Emotion Machine. Simon and Shuster

Mountcastle , V. B. (1997). The columnar organization of the neocortex, Brain, Vol. 20 #4, pp701–722, Oxford University Press, April 1997

Safavi, S., Kapoor, V., Logothetis, N. K., Panagiotaropoulos, T. I. (2014). Is the frontal lobe involved in conscious perception? Frontiers in Psychology 2014; 5: 1063.

Solomonoff, R. J. (1964). A formal theory of inductive inference. Information and Control, 7:1--22, 224—254

Tegmark, M. (2014). Consciousness as a State of Matter. arxiv.org/abs/1401.1219

Tononi, G. (2012). The integrated information theory of consciousness: an updated account. Arch. Ital. Biol. 150, 56–90 (2012

Tononi, G., Boly, M., Massimini, M., Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neuroscience. 17 (7): 450–461

van Gelder, T. (1998). The dynamical hypothesis in cognitive science. Behavioral and Brain Sciences 21, 615–665

Zadra, A., Desautels, A., Petit, D., Montplaisir, J. (2013). Somnambulism: clinical aspects and pathophysiological hypotheses. The Lancet Neurology, Volume 12 , Issue 3 , 285 - 294